Ankita Bhamidimarri 2211CS010025 S3-03

Employee Data Analysis Report

1. Dataset Description

1.1 Source: Formula 1 dataset (multiple CSVs: races, results, drivers, constructors, circuits).

1.2 Columns (example from results DataFrame):

- resultId (int) – unique result ID
- raceId (int) – identifier linking to race
- driverId (int) – identifier linking to driver
- constructorId (int) – identifier linking to constructor/team
- grid (int) – starting grid position
- position (string) – final race position
- points (double) – points awarded
- laps (int) – number of laps completed
- time / milliseconds – finishing time
- fastestLap, fastestLapTime, fastestLapSpeed – performance metrics
- year, round, circuitId, name, location, country – race metadata
- driverRef, surname, dob, nationality – driver attributes

1.3 Data Quality:

- Schema inferred correctly; multiple joins performed between races, results, drivers, constructors, and circuits.
- No critical null-value handling performed; some missing times/statuses exist for DNFs.
- Data types consistent with intended use (IDs categorical, times numeric or string, points double).

2. Operations Performed

2.1 Data Cleaning & Exploration

- Loaded multiple CSV files into Spark DataFrames.
- Inspected schemas and previewed sample rows.
- Performed joins between results, drivers, constructors, and races.
- Selected relevant columns for analysis (driver performance, constructor comparisons, race metadata).

2.2 Descriptive Analytics

- Count of distinct drivers, races, and constructors.
- Distribution of points across drivers and constructors.
- Seasonal performance tracking using year and points.

2.3 Relationship Analysis

- Identified top drivers by cumulative points.
- Visualized constructor dominance (e.g., Ferrari, Mercedes, Red Bull).
- Correlated grid position with finishing position.
- Fastest laps and average speeds analyzed to compare drivers across seasons.

3. Key Insights

3.1 Driver Performance

- Certain drivers consistently dominate (e.g., Lewis Hamilton, Sebastian Vettel, Max Verstappen) across multiple seasons.
- Strong correlation between starting grid and final position — pole sitters often convert to race wins.

3.2 Constructor Insights

- Ferrari and Mercedes show long periods of dominance, with Red Bull emerging strongly in recent years.
- Constructors with stable driver lineups generally outperform those with frequent changes.

3.3 Race Dynamics

- Attrition (DNFs) significantly impacts points distribution — reliability is as important as speed.
- Circuits with higher laps (e.g., Monaco, Singapore) tend to have fewer DNFs but higher variance in position changes.

4. Recommendations

4.1 Strategy for Teams

- Invest in qualifying performance: higher grid positions strongly influence final results.
- Reliability focus: reducing DNFs yields major cumulative point advantages.

4.2 Data Expansion

- Incorporate weather, pit stop, and tyre data to build richer predictive models.
- Track driver age, experience, and career trajectory for performance forecasting.

4.3 Predictive Analytics

- Future work can apply ML models (classification/regression) to predict race outcomes, attrition risk, or championship likelihood.